DOCUMENT RESUME

ED 248 890 IR 011 366

AUTHOR	Wise, Lauress L.; And Others
TITLE	Development and Validation of Army Selection and Classification Measures. Project A: Longitudinal Research Database Plan. Research Report 1356.
INSTITUTION	Human Resources Research Organization, Alexandria, Va.
SPONS AGENCY	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
PUB DATE	Dec 83
CONTRACT	MDA903-82-C-0531
NOTE	99p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Computer Software; *Databases; Data Collection; Data Processing; Editing; *Information Retrieval; *Information Storage; Military Organizations; *Military Personnel; Performance Factors; Performance Tests; *Predictor Variables
IDENTIFIERS	*Army; Computer Security

ABSTRACT

This report describes plans for the development of a major longitudinal research database designed to support the development and validation of new predictors of Army performance, and also new measures of Army performance against which the new predictors can be validated. The following aspects of the database are addressed: (1) the anticipated data elements through Fiscal Year (FY) 1986/87 (the database will include already existing data on FY 81/82 enlistees and their training and scores on the Skills Qualification Test from the Army Enlisted Masterfile, as well as initial predictor measures, training measures, first and second tour performance measures, an experimental predictor battery, and Army-wide performance measures); (2) editing procedures for assuring the accuracy of the data entered, such as linkage to other files, elimination of duplicate records, individual field edits, and machine correction or imputation; (3) storage and access procedures, including the anticipated structure for primary soldier, applicant, sample soldier, soldier progress, field test, and task files, as well as updating procedures; (4) documentation and dissemination formats and standards; and (5) security procedures. (Author/LMM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED248890

Research Report 1356

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Development and Validation of Army Selection and Classification Measures

Project A: Longitudinal Research Database Plan

Lauress L. Wise and Ming-mei Wang
American Institutes for Research

Paul G. Rossmeissl
Army Research Institute

Selection and Classification Technical Area
Manpower and Personnel Research Laboratory

IR011366

U. S. Army
Research Institute for the Behavioral and Social Sciences

December 1983

Approved for public release; distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

L. NEALE COSBY
Colonel, IN
Commander

Research accomplished under contract
to the Department of the Army

Human Resources Research Organization

Technical review by

Hilda Wing
Rebecca Oxford-Carpenter
Michael G. Rumsey

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 1356
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): DEVELOPMENT AND VALIDATION OF ARMY SELECTION AND CLASSIFICATION MEASURES
PROJECT A: LONGITUDINAL RESEARCH DATABASE PLAN

5. TYPE OF REPORT & PERIOD COVERED: October 1982 - September 1989

6. PERFORMING ORG. REPORT NUMBER: FP-PRD-83-7

7. AUTHOR(s): Lauress L. Wise, Ming-mei Wang (American Institutes for Research)
Paul Rossmeissl (U.S. Army Research Institute)

8. CONTRACT OR GRANT NUMBER(s): MDA903-82-C-0531

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Human Resources Research Organization
1100 South Washington Street
Alexandria, VA 22314

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q263731A792

11. CONTROLLING OFFICE NAME AND ADDRESS:
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, VA 22333

12. REPORT DATE: December 1983

13. NUMBER OF PAGES: 98

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): --

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

The Army Research Institute technical point of contact is Dr. Paul G. Rossmeissl. His telephone number is (202) 274-8275.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Database
Data editing
Data security
Data documentation
Predictor measures
Criterion measures
Performance measures

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This research report describes plans for the development of a major longitudinal research database. The objective of this database is to support the development and validation of new predictors of Army performance and also new measures of Army performance against which the new predictors can be validated. This report describes the anticipated contents of the database, editing procedures for assuring the accuracy of the data entered, storage and access procedures, documentation and dissemination procedures, and database security procedures.

Research Report 1356

Development and Validation of Army Selection and Classification Measures

Project A: Longitudinal Research Database Plan

Lauress L. Wise and Ming-mei Wang
American Institutes for Research

Paul G. Rossmeissl
Army Research Institute

Submitted by
Newell K. Eaton, Chief
Selection and Classification Technical Area

Approved as technically adequate
and submitted for publication by
Joyce L. Shields, Director
Manpower and Personnel
Research Laboratory

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

December 1983

Army Project Number
2Q263731A792

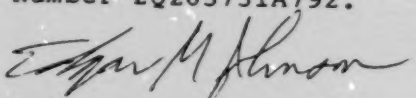
Manpower and Personnel

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

This document describes a path for a key element of a long range research effort for improving the selection, classification, and utilization of Army enlisted personnel. The thrust for this effort came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB) and other selection variables as predictors of training and on-the-job performance. The portion of this effort described herein is devoted to the development of a longitudinal research data base (LRDB) to support the research being undertaken by ARI in-house and through two major contracts: the first being devoted to the development and validation of Army selection and classification procedures (Project A), and the second (Project B) devoted to the development of a Prototype Computerized Personnel Allocation System (EPAS). Together these Army Research Institute efforts, with their in-house and contract components, comprise a landmark program to develop a state-of-the-art, empirically validated personnel selection, classification, and allocation system. The work towards the development of the LRDB is funded primarily by Army Project Number 2Q263731A792.



EDGAR M. JOHNSON
Technical Director

TABLE OF CONTENTS

		Page
1.	INTRODUCTION	
1.1.	Nature and Purpose of the Overall Project	1
1.2.	Role of the Longitudinal Research Database (LRDB)	6
1.3.	Overview of LRDB Contents	7
1.4.	Specific Objectives of the LRDB Plan	11
2.	CONTENTS OF THE LRDB	13
2.1.	Data Elements for the 81/82 Cohort	15
2.1.1.	81/82 APPLICATION AND ACCESSION DATA	15
2.1.2.	TRAINING DATA	22
2.1.3.	DATA FROM THE ENLISTED MASTERFILE (EMF)	24
2.1.4.	SKILLS QUALIFYING TEST (SQT) DATA	32
2.2.	FY83/84 Cohort Data	33
2.2.1.	INITIAL PREDICTOR DATA	33
2.2.2.	TRAINING MEASURES	37
2.2.3.	FIRST TOUR PERFORMANCE MEASURES	39
2.2.4.	SECOND TOUR PERFORMANCE MEASURES	40
2.3.	FY86/87 Cohort Data	40
2.3.1.	EXPERIMENTAL PREDICTOR BATTERY	40
2.3.2.	TRAINING DATA	41
2.3.3.	ARMY-WIDE PERFORMANCE MEASURES	41
2.3.4.	SECOND TOUR DATA	42
3.	EDITING SPECIFICATIONS	43
3.1.	Editing Specifications for FY1981/82 Cohort Training Data	43
3.1.1.	LINKAGE TO OTHER FILES	43
3.1.2.	ELIMINATION OF DUPLICATE RECORDS	45
3.1.3.	INDIVIDUAL FIELD EDITS	47
3.1.4.	MACHINE CORRECTION OR IMPUTATION	50
3.2.	Editing Other FY81/82 Data	51

		Page
4.	DATABASE STORAGE AND ACCESS PROCEDURES	52
4.1.	The Use of RAPID	52
4.2.	Anticipated File Structure	53
4.2.1.	PRIMARY SOLDIER FILES.	56
4.2.2.	APPLICANT FILES.	56
4.2.3.	SAMPLE SOLDIER FILES	57
4.2.4.	SOLDIER PROGRESS FILE.	58
4.2.5.	FIELD TEST FILES	59
4.2.6.	MOS FILES.	59
4.2.7.	TASK FILES	60
4.3.	Updating Procedures.	60
4.3.1.	IDENTIFICATION AND ACQUISITION OF NEW DATA AND RELATED DOCUMENTATION.	61
4.3.2.	LINKING IN RELATED DATA.	62
4.3.3.	EDITING.	63
4.3.4.	DOCUMENTATION.	64
4.3.5.	MERGING THE NEW DATA	64
4.3.6.	DISSEMINATION.	64
4.3.7.	EXCEPTIONS	65
4.4.	Access	65
5.	DATA DOCUMENTATION AND DISSEMINATION	68
5.1.	Documentation Formats and Standards.	68
5.1.1.	EVENT FILE	70
5.1.2.	INSTRUMENT FILE.	71
5.1.3.	SAMPLE STRUCTURE FILE.	72
5.1.4.	DATASET LOGS	73
5.1.5.	CODEBOOKS.	73
5.1.6.	VARIABLE CROSS-REFERENCE FILE.	74
5.1.7.	DATA HISTORY DOCUMENTATION	75
5.2.	Dissemination.	76
6.	DATABASE SECURITY.	77
6.1.	The Need for Security.	77
6.2.	Security Procedures.	78
6.2.1.	SSN ENCRYPTION	78
6.2.2.	CONTROLLED FILE ACCESS	79
6.2.3.	LRDB LOG	83
6.2.4.	OTHER PHYSICAL SECURITY PROCEDURES	84

Page

6.2.5. DATA ENTRY AND EDITING SECURITY PROCEDURES. 85
6.2.6. DATA ANALYSE3 AND REPORTING SECURITY PROCEDURES. 88

6.3. Summary. 89

LIST OF TABLES

Table

1 LRDB File Designators 55

1. INTRODUCTION

1.1. Nature and Purpose of the Overall Project

The Army Research Institute (ARI) is currently funding two large-scale research projects in order to develop a new selection and classification system that will improve the efficiency of personnel utilization in the Army. The purposes of the first project, Project A: Development of Improved Army Selection and Classification Systems, are:

(1) to validate current predictors (primarily the ASVAB composites, supplemented by other available predictors such as high school graduation, Military Applicant Profile (MAP) data, and physical capacities);

(2) to develop new or improved predictors and performance measures; and

(3) to conduct a longitudinal validation of current and newly developed classification measures for prediction of the enlistee's performance from training through the second tour of duty.

The second project, Project B: Development of a Enlisted Personnel Allocation System (EPAS), is to develop a state-of-the-art system, for implementation at the Military Entrance Processing Station (MEPS), to facilitate the initial enlistment decisions. The success of the EPAS depends on a set of cost-effective classification measures that can accurately predict the recruit's future performance throughout his/her Army career. The major objective of Project A is to ensure that such a set of predictors is available to drive the new allocation system. Thus, although the two projects are being conducted separately, they have a common goal to improve the Army's personnel decision mechanism and thereby increase the overall effectiveness of the Army.

The selection of a set of cost-effective classification measures for use in the EPAS requires first a careful evaluation of the relationships of the current predictors to performance on Army jobs. The Army presently has a systematic testing and validation program to support its current selection practice. Specifically, the Armed Forces Vocational Aptitude Battery (ASVAB) is administered to all applicants, and aptitude area composites are developed for use in the initial selection and assignment to training courses. The ASVAB composites are traditionally validated in terms of the extent to which

they predict the enlistee's success in training (essentially measured by course grades). Although these composites generally are quite effective in predicting how well the enlistees will perform during training, their validities for predicting other important areas of Army performances -- general soldiering and job-specific performances -- have not been extensively investigated because valid, sound, and economical measures of these additional aspects of performance are not currently available.

In addition to valid predictions of performance, the new EPAS will require information on the relative utility to the Army of different levels of performance in different MOS. The collection and analysis of such data is another major objective of Project A.

While the greatest concern is with initial selection and classification decisions, Project A will also address subsequent personnel allocation decisions. Two major decision points will be investigated. The first point is the decision at the end of training whether to pass a recruit out of training, or to recycle him or her into additional training for the same or some other MOS, or to drop him or her from the Army altogther. At this point, both pre-induction and training performance measures will

be used to predict subsequent performance in the MOS. The second major decision point is at the completion of each soldier's first tour. At this point the Army must decide whether to encourage or to bar the soldier's reenlistment for a second tour. Here first-tour performance measures must be used, along with the training and pre-induction measures, to predict second-tour performance. Thus, valid training and first-tour performance measures are needed both as criteria for the validation of earlier prediction measures and as predictors of subsequent criteria.

In order to accomplish these objectives, the project is organized into five major tasks. These are:

Task 1: Validation

Task 2: Prediction of Job Performance

Task 3: Measurement of School/Training Success

Task 4: Assessment of Army-wide Performance

Task 5: Development of MOS-Specific Performance Measures

In the course of developing these new or improved measures, there will be pilot field tests in order to assess the psychometric characteristics of the measuring instruments and the ease of their administration. Based on

these pilot tests, the instruments may be revised and then employed in a large-scale field test to collect data for two purposes. The first purpose will be to evaluate formally the effectiveness of the experimental set of pre-induction predictors for predicting success in the Army. The second purpose will be to examine the practical value of using early performance measures as additional predictors of later performances. The results of these evaluations will be employed to guide further refinements of the experimental measures. Through this iterative development and refinement process, a final set of predictor and performance measures will be selected and administered to a cohort of enlisted personnel. The data collected in these administrations will be analyzed for the validation of the classification battery to be employed in the Army's new selection and classification system.

In conjunction with this complete, longitudinal validation of the classification measures, prediction models of enlistees' performance in the Army will be obtained to generate numerical inputs into the EPAS for determining optimal person-job matches. The development of a dynamic allocation procedure will be most useful to the Army if it uses information accumulated through the early period of the enlistee's career to modify post-enlistment decisions at various choice points (post-training

reassignment, promotion, and reenlistment). The capability of the EPAS can be enhanced by incorporating such a dynamic decision process. EPAS would then be used to assist in personnel decisions beyond the initial selection and training assignment made at the MEPS level.

1.2. Role of the Longitudinal Research Database (LRDB)

Clearly, the research process of Project A will generate a large amount of interrelated data that must be assembled into an integrated data base that can be accessed easily by the research teams for various analytical purposes. Therefore, one of the major tasks in Project A is to establish and maintain the longitudinal research data base (LRDB). This data base will link together data on diverse measures gathered in the various tasks of Project A and, in addition, incorporate existing data that are routinely collected by the Army. Such a comprehensive LRDB will enable Project A to conduct a full analysis of how information gathered at each stage of the enlistee's progress through his/her Army career can add to the accuracy of predicting later perfomances.

The richness of the LRDB to be created for the project will not only facilitate efficient validation analyses that concern Project A, but will also enable Project B to test

and revise the prototype selection/allocation system. Specifically, Project B will employ data on the training time, subsequent performance, and the utility of subsequent performance levels to develop the classification model and estimate required parameters.

Indeed, the function of the LRDB extends beyond Projects A and B. The data base can also support other research work to be conducted by the ARI staff tc address specific policy issues that may arise.

1.3. Overview of LRDB Contents

In accordance with the Project A Research Plan, three major sets of data will be assembled within the LRDB. The first set will consist of already existing data on FY81/82 accessions. These data will include accession information (demographic/biographical data, test scores, and enlistment options), training success measures, measures of progress or attrition taken from the Enlisted Masterfile (EMF), and specific information on Skills Qualification Test (SQT) scores. This first set of data will be employed to validate the current version of the Armed Services Vocational Aptitude Battery (ASVAB) insofar as that can be done with available criteria. (This cohort was the first to receive Forms 8, 9, and 10 of the ASVAB.)

Recommendations will be made for revisions in the ASVAB Area Composite scores to be used in classification decisions until EPAS becomes fully functional. These analyses of the existing battery will also produce recommendations for needed additions to the predictor battery. Furthermore, the investigation of methodological and conceptual issues that have plagued personnel decision research will begin with the data on this cohort so that practical solutions may be devised for the validation analyses on two subsequent cohorts. (See Task 1 of the Project A Research Plan.)

The second and third sets of data to be assembled into the LRDB will involve substantial new data collection efforts, in addition to the routinely collected data described above. The second set of data will consist of longitudinal information on FY83/84 inductees. This information will be acquired in three data collection phases:

(1) Beginning in the summer of 1983 and continuing until summer of 1984, samples of recruits will be administered a preliminary predictor battery (consisting of available tests that are not currently employed by the Army but that have potential value for predicting performance on

Army jobs). These data will be analyzed to determine the incremental validity of the new tests (tests such as vocational interest and motivation scales) over the existing predictors (in essence, the ASVAB scores). The results of this evaluation will help guide the development of new preinduction predictors. (They will be developed to be parallel to the preliminary predictors that are found to be effective for predicting performance.)

(2) Later in their first tour (June 1985 to October 1985), data on a revised set of predictors, job knowledge tests, and Army-wide and MOS-specific performance measures under development by Project A will be obtained from this sample. These data, together with the existing data on current predictors and school performance indicators, will be employed to conduct a concurrent validation of the initial predictors using both school measures and subsequent performance as criteria. The school measures will also be validated as predictors of subsequent performance. The findings from this concurrent validation analysis will provide the basis for revision and improvement of new instruments and

for choosing the most cost-effective of them for inclusion in the final set of classification measures.

(3) For members of this sample who reenlist, Army-wide and MOS-specific performance measures will be collected during their second tour (June 1988 to September 1988) and merged with existing EMF measures. These data will be used to validate the pre-induction selection measures and early performance in the Army as predictors of second-tour performance.

Once the new measures are refined on the basis of the analyses of the FY83/84 cohort data, they will be administered to a new cohort (FY86/87 inductees) to allow a complete, longitudinal validation of the final classification battery. The data for this final validation will also be collected in three phases. Briefly, the three data collection points are:

(1) From March 1986 until February 1987, samples of recruits from the 19-focal MOS will be tested with the revised predictor battery, and their school data will be obtained.

(2) During their first term of enlistment (June 1988 to September 1988), the sample will be followed up and the Army-wide and MOS-specific performance measures will be obtained. These data will not only support the predictive validation of the classification measures but will also permit analysis of criterion equivalence between training and first-tour performance and between Army-wide and job-specific performance.

(3) Similarly, during their second tour (January 1991 to March 1991), Army-wide and MOS-specific performance measures will again be obtained from this sample. These data will be used for the longitudinal validation of the predictor measures and the investigation of criterion equivalence between first-tour and second-tour performance measures.

1.4. Specific Objectives of the LRDB Plan

As pointed out in the preceding section, the primary role of the LRDB is to support efficient data analyses as required by the research teams of both Projects A and B. To fulfill this role, the LRDB must be created and maintained in coordination with the data collection

activities and the research process. The data collected throughout the research process of Project A and the data to be acquired from existing Army files must be organized and stored in such a way that they are simple and economical to access. Accordingly, the LRDB plan must meet the following objectives:

(1) To develop systematic and efficient procedures for entering and editing the data.

(2) To establish linkages of data from various sources and resolve all data inconsistencies.

(3) To develop and maintain complete documentation of the data organization and contents.

(4) To store both the data and the documentation cost-effectively and to provide fast and easy access to both simultaneously.

(5) To insure the security and integrity of the data.

2. CONTENTS OF THE LRDB

The adequacy of the LRDB depends heavily on the specific variables that are included. It is not sufficient, for example, to specify that "training measures" will be included. The planning and conduct of the validation analyses require knowledge of the specific predictor and criterion measures that will be available and of the data elements that provide qualifying information. Unfortunately, the degree to which specific variables can be listed varies widely for the three main cohorts. For the FY81/82 cohort, most of the specific variables that will be available are now known. For the FY83/84 and FY86/87 cohorts, a great deal of work will go into defining and collecting new items of information. It is not now possible to give a complete list of specifics at this time.

What follows is a listing of specific data elements to be included in the data base. This list should be helpful to all project staff in planning both analyses and future data collections.

Variable names. Before proceeding, however, a word is in order on conventions and standards regarding variable names. There is a wide range of possible naming conventions, ranging from a strict numbering (e.g., VAR129)

to acronym conventions (e.g., HTIN-height in inches) to single word descriptors (e.g., HEIGHT). The naming convention to be used in this project will combine (a) a two character prefix indicating data source with (b) a descriptive label of up to six characters. (Note that eight characters is the maximum variable name length in most statistical packages.)

The two characters indicating data source will consist of an initial character, designating the type of data, followed by a sequencing character (1 through 9 and then A through Z). The type of data codes are:

A - existing applicant/accession data (including existing predictors)

B - new predictor-battery data

E - Enlisted Masterfile data (including existing Army-wide performance measures)

G - new general (Army-wide) performance data

P - existing MOS-specific performance data (e.g., SQT)

M - new MOS-specific performance data

T - existing school/training data

S - new school/training data

Additional codes may be defined for derived variables that combine data from different sources.

In cases where data are extracted from existing datafiles, the established variable name will be used as the descriptive portion of the variable name in our system (in characters 3 through 8). EMF variable names, for example, are limited to five characters and thus fit nicely into the system. In other cases, it may be necessary to shorten the name to six characters. Where appropriate, more obvious mnemonics may replace the variable name in the original file.

2.1. Data Elements for the 81/82 Cohort

2.1.1. 81/82 APPLICATION AND ACCESSION DATA

The Army collects a great deal of information on each soldier at the time that he/she applies to and is accepted into the Army. Some of this information is retained in each soldier's permanent computerized records (the Enlisted Masterfile), but much is not. Some information, such as responses to individual ASVAB items, is not retained in machine-readable form at all.

For the most part, analyses of the FY81-82 cohort will be limited to data that are already in machine-readable form. One exception will be information used in the

Military Applicant Profile (MAP) score. The MAP is a battery of behavioral indicators now administered to all applicants who have not completed high school. It has been discovered that the overall profile score has not been included in the computerized accession file. The answer sheets, including responses to the approximately 60 (noncognitive) items that comprise the profile, have been retained at ARI and are available for entry. We plan to enter this information for a sample of about 2,500 applicants to allow for analyses of the current MAP items.

The following items of information will be taken from the existing accessions file and retained in the data base:

2.1.1.1. Basic Identifying Information

The data will be used to link the accession data to the EMF data. The linking variables will NOT be stored on the main data files: only a scrambled identifier will be retained on the main data files for linking new data. The data needed for linking and checking the validity of the linkage are:

A1SSN	SOC.SEC.NO.
A1NAME4	4 CHAR NAME ABBREVIATION
A1NAME	FULL NAME
A1DOB	DAY OF BIRTH

2.1.1.2. Individual Background Data

These data will be used to identify differences in backgrounds that may be predictive of performance differences. In addition, certain variables may be valuable as moderators, predicting differences in the relationships between predictor and criterion variables. Other variables such as ZIP codes, will permit links to other (e.g., Census) data files containing information on local geographic or economic conditions.

A1HOMADD	STATE/COUNTY CODE OF HOME ADDRESS
A1HOMZIP	HOME ZIP CODE
A1PRSADD	STATE/COUNTY CODE OF PRESENT ADDRESS
A1PRSZIP	ZIP CODE OF PRESENT ADDRESS
A1MARST	MARITAL STATUS
A1NRDEP	NUMBER OF DEPENDENTS
A1YOB	YEAR OF BIRTH
A1MOB	MONTH OF BIRTH
A1CITIZ	CITIZENSHIP
A1SEX	SEX
A1RACE	POPULATION GROUP
A1ETHNIC	ETHNIC GROUP
A1RELIG	RELIGIOUS PREFERENCE

A1HGT	HEIGHT
A1WGT	WEIGHT
A1PULHES	PHYSICAL PROFILE
A1EDYRS	YEARS OF EDUCATION
A1EDCERT	EDUCATION CERTIFICATION

2.1.1.3. Enlistment Information

These data describe the timing and conditions of enlistment. This information is of primary importance in the development of forecasting models by Project B. In addition, some of these variables (e.g., entry date, primary MOS, pay grade) provide the starting points against which the relationship between the test scores and progress will be charted. Other variables (e.g., moral waivers, additional skill indicators) will be useful as additional predictors.

A1ENTDTE	ENTRY DATE
A1PADDTE	PROJECTED ACT. DUTY DATE
A1AITGRD	AIT GRADUATION DATE
A1AAAS	ACCESSION TO ACTIVE ARMY STRENGTH
A1UPSTAT	STATUS CODE
A1ACTDTE	DATE OF ACTION
A1SERV	BRANCH OF SERVICE

A1PRMMOS	PRIMARY MOS
A1TRNMOS	TRAINING MOS
A1ENLOP	ENLISTMENT OPTION GUARANTEED
A1ENOPT1	ENLISTED OPTION
A1DESGOP	DESIGNATED OPTION
A1ENTPRG	PROGRAM FOR WHICH ENLISTED
A1ENLTRM	TERM OF ENLISTMENT
A1ENTST	ENTRY STATUS
A1BONLVL	ENLISTMENT BONUS LEVEL
A1ABGRD	ABBREVIATED GRADE CODE
A1PAYGRD	PAY GRADE
A1GRDDTE	DATE OF GRADE
A1ASI	ADDITIONAL SKILL INDICATOR
A1WAIVER	WAIVER TYPE
A1MORWVR	REASON FOR MORAL WAIVER
A1WAPLVL	WAIVER APPROVAL LEVEL
A1AFEES	AFEES IDENTIFICATION

2.1.1.4. Prior Military Experience

A1YTHPRG	YOUTH PROGRAM
A1CONDBY	YOUTH PROG.CONDUCTED BY
A1YRSCMP	NO.OF YEARS COMPLETED IN YOUTH PROG
A1PRISRV	PRIOR SERVICE
A1SRVBRK	BREAK IN PRIOR SERVICE

2.1.1.5. Delayed Entry Program (DEP) Information

A1DEPDTE	DEP DATE
A1ENTDT2	DATE OF CONTRACT/ENTRY
A1DISDTE	ENTRY OR DISCHARGE DATE
A1ENTRST	ENTRY STATUS
A1DEPPG	PROGRAM FOR WHICH ENLISTED
A1DESDEP	DESIGNATED OPTION
A1OPTDEP	ENLISTMENT OPTION
A1TNGMOS	TRAINING MOS
A1NODEPR	DEP NON-ENLISTMENT REASON

2.1.1.6. Hometown Recruiter Aide Program (HRAP) Information

A1HRAP	HOMETOWN RECRUITER AIDE
A1HRAPLC	HRAP LOCATION

2.1.1.7. Testing Information

A1CYCL	DATE OF CYCLE NUMBER
A1MCAT	MENTAL CATEGORY
A1TSITE	TEST SITE
A1TSESS	TEST SESSION
A1ASVBFM	ASVAB FORM
A1AFQTPC	AFQT PERCENTILE
A1ASVBxx	ALL ASVAB SUBTEST SCORES

Current ASVAB Area Composite Scores

A1ASCVGT	GENERAL TECHNICAL
A1ASCVGM	GENERAL MAINTENANCE
A1ASCVEL	ELECTRONICS
A1ASCVCL	CLERICAL
A1ASCVMM	MECHANICAL MAINTENANCE
A1ASCVSC	SURVEILLANCE/COMMUNICATIONS
A1ASCVCO	COMBAT
A1ASCVFA	FIELD ARTILLERY
A1ASCVOF	OPERATIONS/FOOD
A1ASCVST	SKILLED TECHNICAL
A1ASCVWS	AFWST(WOMEN ONLY)

Previous ASVAB Subtest and Composite Scores

A1PASVFM	PREVIOUS ASVAB TEST FORM
A1PAFQTS	PREVIOUS ASVAB TEST-AFQT
A1PASVxx	PREVIOUS ASVAB SUBTEST SCORES
A1PASCXX	PREVIOUS ASVAB CONPOSITE SCORES

2.1.2. TRAINING DATA

ARI has expended considerable effort to collect training information on 1981 and some 1982 accessions. These data indicate the timing and duration of training, the course(s) taken, the overall outcome, and some measure of performance in the course. It is important to note that the nature of these performance measures varies widely by school and sometimes by course or class within school.

2.1.2.1. Basic Identifying Information

These data will be used for identification purposes only and will NOT be stored on the main data files; only a scrambled identifier will be retained in the main data files for linking in new information.

T1NAME5	5 CHAR ABBREVIATION FOR NAME
T1SSN	SOCIAL SECURITY NUMBER

2.1.2.2. School Identification Information

These data will be used to identify the school, the class, and the course for which the scores on each file have been collected.

T1SCHOOL	SCHOOL/ATC CODE
T1COURSE	NAME OF COURSE
T1CLASS	CLASS ID NUMBER WITHIN COURSE
T1MOSAWD	MOS AWARDED UPON COMPLETION
T1SKLLVL	MOS SKILL LEVEL AWARDED

2.1.2.3. Students' Progress Through the Training Program

Essential to this project are the data which describe each student's progress through training.

T1ENRDTE	ENROLLMENT DATE
T1GRDDTE	DATE OF RECYCLE, TRANSFER, OR GRADUATION
T1ATTRIT	TYPE OF ATTRITION
T1DISP	DISPOSITION (PASS, RECYCLE, TRANSFER OR DROP)
T1SCORE1	STUDENT'S COURSE GRADE OR TEST SCORE
T1STYPE1	TYPE OF SCORE
T1SCORE2	SECONDARY PERFORMANCE MEASURE (FOR SOME MOS)
T1STYPE2	TYPE OF ADDITIONAL SCORE
T1SELECT	WAS SPECIFIC MOS GUARANTEED FOR BASIC INFANTRYMEN
T1MORSE	MORSE CODE TAKEN FOR 05B AND 05C

2.1.3. DATA FROM THE ENLISTED MASTERFILE (EMF)

The Army Enlisted Masterfile (EMF) contains a significant amount of information that is essential to Project A. In particular, information on each soldier's progress through his or her Army career is captured by the EMF. The EMF also contains important information on the individual background and enlistment conditions of each soldier that are important checks against similar information obtained from the Accession files.

While some of the analyses will focus on a specific MOS, others will require information on a broadly representative cohort of soldiers. In particular, in analyzing the generalizability of results from samples of MOS, it is essential that such representative cohorts be analyzed. The EMF provides one source of information on the progress of all recruits, against which the results for specific samples can be compared.

The following list indicates the EMF data elements that will be needed by this project in order to avoid large and redundant data collection costs. The variables are grouped into seven types of information, and the use of each type of information in the planned analyses is indicated. Basic and background information will be

retained only once in the system. Other information, on progress and problems, will be obtained at regular intervals and accumulated into the data base.

2.1.3.1. Basic Identifying Information

Again, these data will be used to link new EMF information to data previously collected on the samples of soliders in the project. The linking variables will NOT be stored on the main data files. Only a scrambled identifier will be retained on the data files for linking in new information. The EMF variables needed for linking and for checking the validity of the linkage are:

E1SSN	SOCIAL SECURITY NUMBER
E1SSNPR	PREVIOUS (INCORRECT) SSN
E1SNCTL	SSN CORRECTION DATE

2.1.3.2. Individual Background Data

These data will be used to identify differences in backgrounds that may be predictive of performance differences. In addition, certain key variables will be used to check the "cultural fairness" of any proposed selection and classification algorithm. Note that most of this information will also be obtained from the accession

files. The corresponding EMF variables will be used to check or verify the accession data. After completing this check, only one copy of this information will be retained. The background data elements from the EMF that are needed to either add or verify essential information include:

E1SEX	SEX
E1RACE	POPULATION GROUP
E1REDCAT	RACIAL/ETHNIC DESCENT
E1ETHNIC	ETHNIC GROUP DESIGNATION
E1CLANG	LANGUAGE IDENTITY
E1CITIZ	CITIZENSHIP STATUS
E1DOB	DATE OF BIRTH
E1MARST	MARITAL STATUS
E1NRDEP	NUMBER OF DEPENDENTS
E1CIVED	ACADEMIC EDUCATION LEVEL
E1MADCD	COLLEGE MAJOR
E1STPD	STATE OF RESIDENCE AT ENLISTMENT

2.1.3.3. Enlistment Conditions

The enlistment data required by this project include physical and mental test scores and information on the terms or conditions of enlistment. The test scores are the primary predictive measures currently available. The

information on enlistment conditions is essential to understanding the relationship between the test scores and subsequent performance in the Army. As with background data, much of the enlistment information will also be obtained from accession files. Again, only one copy of this information will be retained after any inconsistencies are resolved. The required enlistment variables include:

E1ASVBXX	ALL ASVAB AREA COMPOSITE SCORES
E1AFQSC	ARMED FORCES QUALIFICATION TEST SCORE
E1AFQG	AFQT GROUP
E1DLAB	DEFENSE LANGUAGE BATTERY SCORE
E1PHYPR	PHYSICAL PROFILE
E1PHYCA	PHYSICAL LIMITATION CATEGORY
E1XFACT	WEIGHT-LIFTING CAPACITY
E1COMPT	SERVICE COMPONENT
E1ENLOP	ENLISTMENT OPTION CODE
E1MORWA	ENLISTED/REENLISTMENT WAIVER
E1TERMS	TERMS OF SERVICE OR ENLISTMENT
E1BASD	BASIC ACTIVE SERVICE DATE
E1BONIN	BONUS INDICATOR
E1RPFLG	RECRUITER FLAG (PROMOTED OR SEPARATED)
E1RCRCD	RECRUITER CODE
E1PLOEN	STATE OF ENLISTMENT
E1TYPLA	TYPE OF LAST ACCESSION

E1DATLA	DATE OF LAST ACCESSION
E1ETSDT	DATE OF EXPIRATION OF LAST TERM OF SERVICE

2.1.3.4. Basic Progress in the Army

A major outcome to be predicted at the time of selection is the applicant's probable rate of basic progress in the Army. Several EMF variables are needed to chart the progress of the soldiers in the research samples for use in validating new and existing predictor measures. These include:

E1GRTIT	GRADE IN WHICH SERVING
E1DOR	DATE OF RANK
E1PAYGR	PAYGRADE
E1PAYSX	PAYGRADE & SEX
E1GRDDT	DATE OF LAST GRADE CHANGE
E1BPEDT	BASIC PAY ENTRY DATE
E1GRDTT	TYPE OF LAST GRADE CHANGE
E1NCOES	NCO EDUCATION SYSTEM (LEVEL ATTAINED)
E1PROPT	CURRENT PROMOTION POINTS
E1PROPDT	CURRENT PROMOTION POINT DATE
E1PRVPT	PREVIOUS PROMOTION POINTS
E1PRVPDT	PREVIOUS PROMOTION POINTS DATE
E1PROPA	PROFICIENCY PAY STATUS

E1AITDT	AIT GRADUATION DATE
E1PACE	SELF-PACED AIT FLAG
E1EERWA	EER WEIGHTED AVERAGE
E1TUREL	TOUR ELIGIBILITY
E1SECCLR	PERSONNEL SECURITY CLEARANCE
E1SGTID	DRILL SERGEANT QUALIFICATION
E1ADPAY	ELIGIBILITY FOR ADDITIONAL PAY
E1VEAP	VETERANS EDUCATION ASSISTANCE PROGRAM CODE

2.1.3.5. Performance in a Particular MOS

Since much of military performance is specific to particular occupational specialties, many of the criteria used in evaluating new and existing predictor measures will concern progress and performance within an MOS. The specific EMF variables required to track this information are:

E1CMF	CAREER MANAGEMENT FIELD
E1PRMOS	PRIMARY MOS
E1DMOS	DUTY MOS
E1SMOS3	SECONDARY MOS CURRENT (3-POS)
E1PMOTT	TYPE OF LAST PMOS CHANGE
E1PMODT	DATE OF LAST CHANGE TO PMOS
E1PGMOS	PRIMARY PROGRESSION MOS

E1BOMOS	MOS OF BONUS
E1DDSID	ADDITIONAL SKILL INDICATOR, DUTY MOS
E1ADSID2	ADDITIONAL SKILL INDICATOR, PREVIOUS
E1ADSID3	ADDITIONAL SKILL INDICATOR, 2ND PREVIOUS
E1PQDES	PRIMARY MOS, IN WHICH TESTED, SQ DESIGNATOR
E1PQSCR	PRIMARY SQT SCORE (FOR PQDES)
E1PQPFR	SKILL QUALIFICATION PERCENTILE (FOR PQDES)
E1PMOST	PRIMARY MOS IN WHICH TESTED
E1PSQDT	DATE OF LAST CHANGE ON PMOS TESTED (SQT)
E1PMOST1	PRIMARY MOS IN WHICH TESTED, FIRST PRIOR
E1PMOST2	PRIMARY MOS IN WHICH TESTED, SECOND PRIOR
E1PRQDT	DATE OF PREVIOUS CHANGE IN PMOS TESTED
E1PRDES	PREVIOUS PRIMARY MOS IN WHICH TESTED
E1PRQSC	SQT SCORE FOR PREVIOUS MOS (PQDES)
E1PRPER	PREVIOUS SQT PERCENTILE (FOR PQDES)
E1SQDES	SECONDARY MOS SQT
E1SSQDT	SMOS SQT DATE
E1SQSCR	SMOS SQT SCORE

2.1.3.6. Indicators of Attrition and Related Problems

In addition to using measures of Army-wide and MOS-specific progress as criteria, it will also be essential to predict the potential for problems within the Army. In particular, early attrition from the Army for reasons of conduct or low performance represents an outcome that must serve as a negative criterion in the validation of predictor measures. The specific EMF variables required are:

E1CHSEP	CHARACTER OF SEPARATION
E1SPNIS	SEPARATION PROGRAM DESIGNATOR
E1SEPTT	TYPE OF LAST SEPARATION
E1SEPDT	DATE OF LAST SEPARATION
E1DFRDT	DATE OF LAST DROP FROM ROLLS
E1DFRTT	TYPE OF LAST DROP FROM ROLLS
E1STATU	STATUS OF LAST STATUS CODE CHANGE
E1STATT	TYPE OF LAST STATUS CODE CHANGE
E1LAWTT	TYPE OF LAST AWOL TRANSACTION
E1LAWDT	DATE OF LAST AWOL TRANSACTION
E1AWODT	DATE OF RETURN FROM LAST AWOL
E1AWOTT	TYPE OF LAST RETURN FROM AWOL
E1RMCTT	TYPE OF LAST RETURN TO MILITARY CONTROL

2.1.3.7. Reenlistment Eligibility and Conditions

A final indicator of each soldier's value to the Army is whether the soldier is eligible for reenlistment and in fact does reenlist. The specific EMF variables required to capture this information include:

E1EREUP	REENLISTMENT ELIGIBILITY
E1EREUPP	REENLISTMENT ELIGIBILITY BAR
E1VRPMO	SELECTIVE REENLISTMENT BONUS MOS
E1VRMUL	SELECTIVE REENLISTMENT BONUS MULTIPLIER
E1VRGRD	SELECTIVE REENLISTMENT BONUS PAY GRADE
E1VRRDT	ENLISTMENT/REENLISTMENT BONUS DATE
E1VRPNR	ENLISTMENT/REENLISTMENT BONUS PAYMENT NO.
E1VRTRM	ENLISTMENT/REENLISTMENT BONUS PAYMENT TERM
E1PSVCI	NUMBER OF TIMES ENLISTED/REENLISTED

2.1.4. SKILLS QUALIFICATION TEST (SQT) DATA

Special datafiles will be obtained containing SQT score information for soldiers in the FY81/82 accession cohort. These data will significantly expand the SQT information available in the EMF by adding scores on tests not released for operational use and by adding information on the particular form (skill level, test year and track) completed by each soldier. The specific data elements to

be included, apart from identifying information used only for linkage, are:

P1MOS	MOS IN WHICH TESTED
P1SKLLVL	SKILL LEVEL TESTED
P1TRACK	FORM OF SQT AT THIS LEVEL
P1YEAR	SQT TEST YEAR
P1TESTDT	DATE OF TESTING
P1SQTSCR	SQT SCORE

2.2. FY83/84 Cohort Data

The data assembled for the FY83/84 cohort will include all of the data assembled for the FY81/82 cohort from existing sources plus a considerable number of new measures developed by the project. It is not possible to specify the exact variables at this time, but a summary of these new measures is included below.

2.2.1. INITIAL PREDICTOR DATA

All of the application and accession variables collected for the FY81/82 cohort will also be assembled for the FY83/84 cohort. One significant change in these data is that Forms 11, 12, and 13 of the ASVAB will have been introduced. In addition, Task 2 will develop and administer batteries of additional predictor measures.

2.2.1.1. Preliminary Battery

A preliminary battery of predictor measures will be administered to special samples of about 2,100-4,600 FY83/84 accessions from October 1983 to June 1984 in each of four MOS:

MOS	Title	Training Site
05C	Radio TT Operator	Ft. Gordon, GA
19E/K	Tank Crewman	Ft. Knox, KY
63B	Vehicle and Generator Mechanic	Ft. Dix, NJ Ft. Leonard Wood, MO
71L	Administrative Specialist	Ft. Jackson, SC

These data will be collected during the first week of the soldiers' advanced training (AIT). Training school achievement measures (developed in Task 3) will also be collected as enlistees pass through these training courses and will be used as criteria in the initial analysis of Preliminary Battery measures.

This preliminary battery will focus on types of predictors not currently in use. Analysis of these measures will allow an early determination of the major human attributes not assessed by the current pre-induction battery, and whether the measurement of these attributes significantly increases the accuracy with which performance

is predicted. This information will be useful for guiding the development of new predictors into areas most likely to increase the accuracy of prediction and classification.

The Preliminary Battery must necessarily be made up of "off-the-shelf" instruments, because there is too little time prior to the scheduled administration of the Preliminary Battery to develop and pilot test new measures of constructs deemed potentially useful. The testing will probably be done within a four-hour time, since soldier time at AIT schools is generally allocated in four-hour blocks. That time period is sufficient to administer "off-the-shelf" measures of biographical information, vocational interest, motivation, and cognitive ability. Psychomotor measures will probably not be included in the Preliminary Battery because of the time constraints.

2.2.1.2. Trial Predictor Battery

A trial battery of predictor measures, following pilot testing for practice effects, fakeability, and motivational set (with the pilot test administered to samples of the FY81/82 cohort) will be administered to an average of 500 soldiers in each of the 19 MOS. These data will be collected between June and October 1985 from FY83/84 cohort

members who will generally be in the third year of their first term of enlistment. Job-performance criterion data will be collected for these same soldiers for use in a concurrent validation of the Trial Battery. The current plan is to develop a Trial Battery that will require a maximum of four hours to administer, including computer-administered and apparatus measures.

In addition to the collection of these primary data, four research projects will be undertaken. First, to measure test-retest reliability, the predictor battery will be readministered to a subsample of 500 soldiers 30 days after the initial administration. Second, to measure practice effects, a subsample of 115 soldiers will be readministered the battery in the week following the first testing. Third, to measure fakeability, 115 soldiers will be instructed to "fake good" and another 115 soldiers will be instructed to "fake bad" on the non-cognitive portions of the paper-and-pencil battery. Finally, to measure score differences between "early career" soldiers (i.e., new recruits) and the primary sample (later career soldiers) in examining maturational effects, a sample of 1,000 new recruits will receive the battery.

2.2.2. TRAINING MEASURES

Currently available training measures will be obtained from the school records for input into the LRDB. In addition, scores from job knowledge tests, MOS content tests, performance ratings, and end-of-course knowledge tests (EOCKT) will be added to the file.

2.2.2.1. Available Records

Training performance measures that have been identified in Task 3 as adequate indicators, based on interview data and on qualitative analyses, will be obtained from school records for all recruits in the FY83/84 cohort who receive training in one of the 19 MOS selected for this research project. Task 3 staff will arrange for the school to provide the required training data to be input into the LRDB on a continuing basis from July 1983 to September 1984, as each new class completes training.

2.2.2.3. Prototype Measures

Preliminary and revised prototype performance measures will be administered to samples averaging 575 soldiers from four MOS: 05C, 63B, 19E/K, 71L. Different test formats will be examined, including free response measures and synthetic hands-on performance measures. In addition, measures of general performance in training and new indices from existing measures will be obtained for samples from all 10 MOS. The data collected on these prototype measures will be analyzed to determine the relative feasibility and value of the administration of each type of measure.

2.2.2.4. End-of-Course Knowledge Tests (EOCKT)

Revised EOCKT will be gathered on samples averaging 500 soldiers from the 19 MOS. These will be obtained at the same time that the other performance measures are obtained for the FY83/84 cohort.

2.2.3. FIRST TOUR PERFORMANCE MEASURES

Concurrent with the administration of the Trial Predictor battery (during the latter half of 1985, see section 2.2.1.2), Army-wide performance measures will be collected from the same 19 MOS samples. These data will include rating scale measures and behavioral indices generated from records of commendations, disciplinary problems, and attrition. For half of these samples (9 MOS), MOS specific performance measures will also be administered. The tentative list of MOS includes the following:

11B	Infantryman
13B	Cannon Crewman
19E/K	Tank Crewman
05C	Radio TT Operator
63B	Vehicle and Generator Mechanic
64C	Motor Transport Operator
71L	Administrative Specialist
91B	Medical Care Specialist
95B	Military Police

The measures used will include hands-on task performance tests as well as job knowledge tests and supervisor and peer ratings.

2.2.4. SECOND TOUR PERFORMANCE MEASURES

Army-wide and MOS-specific performance measures will also be collected from the FY83/84 cohort during their second tour (June 1988 through September 1988). Samples of about 100 soldiers are expected for each of 10 different MOS (05C, 63B, 71L, 19E/K and 64C, 76Y, 91B, 94B, 11B, 13B) for which first tour MOS-specific performance measures are obtained. The measures will be revised versions of the first tour performance measures.

2.3. FY86/87 Cohort Data

The data collected on the FY86/87 cohort will be parallel to the data collected on the FY83/84 cohort, except that concurrent predictor measures will not be collected on the FY86/87 cohort. Data from existing accession and EMF records will be gathered along with data from the predictor and criterion measures developed by this project.

2.3.1. EXPERIMENTAL PREDICTOR BATTERY

From March 1986 through February 1987, the revised predictor battery will be administered to samples of recruits at the beginning of AIT. Current plans call for

testing an average of 2,200 recruits in each of the 19 focal MOS. (Data will be collected from additional MOS if preliminary analyses indicate that other MOS are required to assure sufficient validity generalization.)

2.3.2. TRAINING DATA

Training performance data will be obtained from schools for the FY86/87 cohort sample who receive training in the 19 focal MOS between March 1986 through May 1987. The measures collected will include the EOCKT as well as those prototypical measures that prove feasible and valid.

2.3.3. ARMY-WIDE PERFORMANCE MEASURES

The Army-wide (Task 4) and the MOS-specific (Task 5) performance measures administered to the FY83/84 cohort will be revised on the basis of analyses of these data. The revised performance measures will be administered to analogous samples of the FY86/87 cohort for use as final validation criteria.

2.3.4. SECOND TOUR DATA

Again, revised Army-wide and MOS-specific performance measures will also be administered to samples of the FY86/87 cohort who remain for a second tour of duty.

3. EDITING SPECIFICATIONS

For each set of data to be entered into the data base, a detailed set of editing specifications will be developed, reviewed, revised, and implemented. These specifications will give procedures for linking the new data to existing records, identifying erroneous or improbable values, correcting these values, and replacing missing values where appropriate. Editing specifications for the FY81/82 cohort training data are given below as an example.

3.1. Editing Specifications for FY81/82 Training Data

3.1.1. LINKAGE TO OTHER FILES

Prior to the detailed editing of each field, the 1981 training data will be linked to the FY81/82 Accession data file and to the 1982 year-end EMF file. The reason for this prior linkage is two-fold. First, the 1981 training data file contains records on some number of soldiers who are not of interest to the current study. These include soldiers not in the regular Army, soldiers who actually entered prior to FY81, and soldiers who are actually reenlistees. By eliminating these soldiers first, editing resources can be concentrated on the cases of primary

interest. The second reason for prior linkage is that the information from the Accession and EMF files will provide important checks on the reasonableness of the training data fields and will provide information essential to the correction of missing or invalid values.

The linkage of additional data will be accomplished in two stages. The first stage will involve matching the training records to a special "Link" file which contains identifying information on the soldiers of interest. (See the discussion of the "Link" file in Section 4.3.) For the training records that match a record in the Link file, identifying information will be stripped and replaced with the scrambled identifier from the Link file. This scrambled identifier serves as the primary key for matching data already in the data base. The second stage will be to merge the training data with other information in the data base using the scrambled identifier.

Two passes will be used in the initial match to the Link file. The initial pass will match on SSN. For each training record which does not match a Link file record, a second match will be attempted. The purpose of this match is to identify errors in the coding of SSNs. In this second pass, the training records will be matched to the

Link file on the basis of the name field (actually the first 5 characters of name) and on MOS. (It is expected that each of the initially unmatched training records may match many Link file records.) For each match, the SSNs will be compared and a new variable, NMATCH, will be computed as the number of matching digits. A frequency distribution will be run on NMATCH to determine an appropriate cutoff point for accepting a match. (We currently expect to accept matches with 7 or more digits in common.) Accepted matches will have all identifying information replaced with the scrambled identifier and will be merged with the main datafile by scrambled identifier. In addition, a dummy record with the alternate SSN will be inserted into the Link file for use in future matches.

3.1.2. ELIMINATION OF DUPLICATE RECORDS

The training datafile is known to contain both exact duplicate records and also valid instances of multiple records for the same soldier due to recycling. The next step in the editing process will be to eliminate the exact duplicates and create a CYCLENO variable (which numbers the training courses taken by an individual soldier sequentially beginning with 1 for the course with the earliest enrollment date) for other instances of multiple

records. The CYCLENO variable, together with the soldier's scrambled SSN, will uniquely identify each valid record in the training file.

The first step in this process is to eliminate all records where the preceding record contained identical values for all fields of interest; in this case, all fields except name. After this has been completed, a second pass will be made to identify obviously valid recycles. The file will be sorted by ID and by T1GRDDTE (graduation/recycle date). The first record for each soldier will have CYCLENO set to 1. Subsequent records will be accepted as valid and have the CYCLENO variable increased by 1 if the following conditions are met:

(1) the disposition variable (T1DISP) in the preceding record has a value of A or B (recycle or transfer)

(2) the T1GRDDTE variable for the current record is at least 10 days greater than the T1GRDDTE value for the preceding record. (A frequency distribution on this difference will be run to check the reasonableness of this cutoff date.)

All duplicates not meeting these two criteria will be sent to an error file, printed, and inspected by hand for further resolution. It is expected that these records will either be true duplicates with data entry errors or valid recycles with errors in T1DISP or T1GRDDTE.

3.1.3. INDIVIDUAL FIELD EDITS

The editing specifications for each field are given below. (See Section 2.1.2. for a list of the variable names to be edited.) In each case, error records are to be listed individually and manually inspected for error resolution. Where a large number of errors occur in a given field, machine corrections will be developed as appropriate. In each case, a default procedure for the imputation of missing or invalid data is given.

a. T1MOSAWD: values must match the list of valid MOS for this field. A cross-tabulation of T1MOSAWD by A1TRNMOS (training MOS from the accession file) is to be run to resolve invalid values. For records where the T1MOSAWD code is invalid or missing and an EMF record has been linked, the E1PMOS and E1DMOS variables will also be used in error resolution.

b. T1SCHOOL: must be a valid school code for this MOS.

c. T1COURSE: must be a valid code for this school and MOS.

d. T1CLASS: must be a valid code for this course and school.

e. T1SKLLVL: must be a valid code for this MOS.

f. T1ENRDTE: must be a valid date, less than T1GRDDTE, and greater than or equal to A1ENTDTE. A distribution will be run on the number of days between A1ENTDTE and T1ENRDTE to establish an appropriate cutoff for the identification of outliers. (Note that this edit may also catch errors in A1ENTDTE.) In most cases, T1ENRDTE must be identical for specific course and class codes. In such cases, the modal value will be substituted for missing or invalid values.

g. T1GRDDTE: must be a valid date and greater than T1ENRDTE. Graduation date values will also be compared with the modal value among graduates of the same course and class. For recycles and attritions, the value must be less or equal to the modal value except in the case of self-paced classes.

h. T1DISP: must be a valid code for this field. A table of T1DISP by T1ATTRIT will be examined to determine valid combinations. Basically, T1ATTRIT should be blank for graduates (F, G, or H) and for progressive transfers (E) and nonblank for recycles, attrition transfers, and relief (A,B, or C).

i. T1ATTRIT: must be a valid code and consistent with T1DISP as specified above. If attrition is indicated prior to 30 September 1982 and an EMF record is matched, the attrition code will be compared to E1CHSEP (character of separation) and E1SEPTT (type of separation) and T1GRDDTE compared to E1SEPDT separation date. Frequencies and cross-tabulations will be run to determine which combinations are to be treated as errors.

j. T1SCORE1: Frequency distributions will be run for each school, MOS awarded and course to determine cutoff values for the identification of outliers. For some MOS, the scores will be compared to the T1DISP and T1ATTRIT values to assure that scores are either missing or below a cutoff value for recycles or academic attritions. Existing documentation and subsequent inquiries will be required to complete the specification of the treatment of the field and to create a type of score value, T1STYPE1, that allows for proper interpretation of this field.

k. T1SCORE2: For certain MOS, a second score was recorded. This field will be created with appropriate analyses of outliers in accordance with existing documentation. A second score type variable, T1STYPE2, will also be created. Two additional variables will be generated from data initially in the T1SCORE2 field. For MOS 05B and 05C, a variable T1MORSE will indicate completion of Morse code training. For MOS 11X, the actual MOS awarded will be determined and a variable, T1SELECT, created to indicate whether the awarded MOS had been originally guaranteed.

3.1.4. MACHINE CORRECTION OR IMPUTATION

After manual inspection of all error records, resolvable cases will be updated and the initial edit will be rerun. For cases where missing or invalid values remain, imputed values will be substituted. (Each variable imputed will also be flagged with a binary flag so that imputed values can be identified and, if desired, deleted in later analyses.)

For the categorical variables, "predictor" variables are already indicated in the above consistency edits. In each case, imputed values will be generated randomly with

probabilities proportional to the conditional distribution of the variable in question (conditioned on the values of the predictor variable(s)). In many cases, this simply means substituting the one school code where this MOS is taught if the school code is missing, or making the course code consistent with the school and MOS codes. In other cases, values may actually be generated probabilistically.

For continuous variables (T1SCORE1 and T1SCORE2), the SAS procedure PROC IMPUTE will be used to generate imputed values from initial ASVAB test scores.

In all cases, the exact details of the machine procedures for error resolution will be refined using information from the outcome of the editing procedures.

3.2 Editing Other FY81/82 Data

The editing of other FY81/82 data (Accession, Applicant, EMF, and SQT) will proceed in a similar fashion. After initial linkage, editing will proceed variable-by-variable, using the best available information to test or correct the data in each field. Copies of the appropriate Army Regulations will be obtained to aid in the editing as well as the documentation of each field.

4. DATABASE STORAGE AND ACCESS PROCEDURES

4.1 The Use of RAPID

At the time that the proposal for this project was developed, RAPID was identified as the most cost-effective data base management system (DBMS) that meets Project A needs. This decision was based on three important features of RAPID. The first was the storage and access mode employed by RAPID. RAPID uses a "transposed file" organization, which means that it stores together all the information on a single variable rather than all of the information on a single "case" or respondent. It stores the data in a direct access file with appropriate indices so that it can read selected variables without having to read through the entire file. The standard statistical packages, in contrast, employ a sequential access mode and store data by case. Even when only a few cases and variables are required, the entire file must be read in order to select the desired information. Most other common DBMSs do use direct access files, but still store information by case so that they only add additional overhead in accessing selected variables.

The second important feature of RAPID is that it provides for a significant degree of data compression. This means that it will be feasible to store much more of the data on mass storage units, greatly increasing the speed with which these data can be retrieved in comparison to tape storage.

The final advantage of RAPID is that it provides convenient interfaces with both SAS and SPSS (as well as other) statistical packages. This facilitates the creation of special analysis files and the use of SAS to manipulate data to be loaded into the data base.

4.2. Anticipated File Structure

RAPID is a "relational" data base system. It processes a series of "relations" which may be viewed as data tables where the columns are different variables and the rows are different observations. Each row is "identified" by one or more columns which provide the keys for accessing the information in the table. Each row must have a unique combination of key values.

Relations are normalized if they contain no "redundant" information. This frequently means creating several subfiles with different fields. In the FY81/82

training file, for example, the MOS, school, course, and class information are constant for all soldiers in the same class. A more efficient storage of information would result from maintaining course and class information in a separate relation (file) with only one entry (row) for each course and class and then keeping only an index to this information on the individual soldier records.

Determining the "optimum" arrangement of data into separate files or relations requires analysis of the trade-off between reduction in data storage requirements and reductions in processing costs, when only the smaller file(s) need to be accessed, and the corresponding increase in processing costs, when it is necessary to join information separated into different files. At present, we can only forecast requirements approximately so an exact optimization is not possible. During the course of the project, statistics on actual access requirements will be used to reevaluate our file and subfile design.

The organization of data into files currently planned is given below along with some discussion of the rationale behind the proposed organization. Table 1 summarizes the different file types that are planned and gives a three-character designator for each type that will be used as a prefix in the file name.

TABLE 1

LRDB FILE DESIGNATORS

PSF - Primary Soldier Files, one file for each cohort, one record for each soldier in the corresponding cohort, keyed by scrambled ID.

APF - Applicant File, one file for each cchort, one record for each application not leading to accession, keyed by scrambled ID and application number.

SSF - Sample Soldier Files, a separate file for each of the MOS selected for special data collection (FY83/84) and FY86/87 cohorts only), keyed by scrambled ID within file.

SPF - Soldier Progress File, one file for each cohort, one record for each EMF record pulled (tentatively 4 EMF records per year) for each individual in the corresponding Primary Soldier File, keyed by scrambled ID and month of enlistment.

FTF - Field Test Files, one file for each field test event, one record for each soldier tested, keyed by scrambled ID.

MOS - MOS Files, one file, one record for each MOS, keyed by MOS.

TSK - MOS/TASK Files, one file, one record per MOS and Task, keyed by MOS and task code.

4.2.1. PRIMARY SOLDIER FILES

There will be three primary soldier files, one for each of the three main cohorts. These files will contain all of the "constant" information on each accession in the cohort (i.e., each accession during the period that defines the cohort). This information will include all information from the current accession record, information on the completion of training, and information on reenlistment decisions. This file will be keyed by soldier identifier (scrambled SSN).

An abbreviated primary soldier file will be maintained for each of the gap periods between the three main cohorts. These files, which will contain only accession information, will be of primary use to Project B in the development of forecasting models.

4.2.2. APPLICANT FILES

A separate applicant file will be maintained for the accession period corresponding to each of the three cohorts. This file will be keyed by the same scrambled identifier used in the Primary Soldier File and by an occurrence number within each individual ID. There will be one record for each application of each individual. In

order to avoid duplication, only application information not leading to an accession of interest will be kept here. By concatenating these files with the Primary Soldier Files, however, a complete set of application data can be obtained.

Each record will contain test scores and other information relating to the particular application including background data that ought not to change from one application to another but might change anyway. These data will be useful in establishing overall base rates for applicants and for looking at the level of consistency in different variables across applications.

4.2.3. SAMPLE SOLDIER FILES

For the FY83/84 and FY86/87 cohorts, there will be Sample Soldier Files consisting of all of the soldiers sampled for special data collection. Currently, we plan to maintain 19 separate files corresponding to the 19 different MOS sampled for new data collection. This will facilitate the creation of separate analysis files for each MOS. It will, of course, be a simple matter to concatenate these files for across-MOS analyses.

The Sample Soldier Files will also be keyed to an alternate identifier defined as the index number of the corresponding Primary Soldier File record. They will not contain any other variables stored in the Primary Soldier Files, but they will be directly linkable to the Primary Soldier Files by this index number (without further sorting). These files will contain all of the new measures collected on each soldier in the selected samples. Some of the measures collected will vary from one MOS to the next, particularly the MOS performance measures and the job knowledge and hands-on measures collected during training. (This is a major reason for maintaining separate files by MOS.) It is likely that these files will also be further divided by data collection period. The FY83/84 second-tour sample, for example, will be only a subset of the concurrent validation samples, and the concurrent validation samples will also be different from the samples receiving the Preliminary Predictor Battery.

4.2.4. SOLDIER PROGRESS FILES

Separate Soldier Progress files will be used to store recurring information on each soldier's progress in the Army. There will be separate Soldier Progress files for each cohort. These files will be keyed by soldier ID and a

generated variable TOURMON which gives the number of months since the beginning of active service. The contents of these files will come primarily from the EMF, which will be accessed at regular intervals, and from special SQT files. The primary purpose of these files is to provide the basis for time series or career trajectory analyses in which each soldier's progress is charted as a function of time in service.

4.2.5. FIELD TEST FILES

A separate datafile will be created for each field test of each new instrument or battery. These files will be keyed by alternate identifier (index number in the relevant Primary Soldier File) so as to be readily linkable to all other information on the same soldiers. The contents of each file will be highly specific to the related field test.

4.2.6. MOS FILES

A separate file will be maintained which will contain information on the characteristics of each MOS. This file will be keyed by the three-character MOS code. The specific contents of the file are not fully known at this

time. Some information on qualifications for each MOS, workforce size and requirement forecasts, training location(s), and utility measure data collection will be included.

4.2.7 TASK FILES

Information on specific tasks performed within each MOS is available from several sources (e.g., the Army Occupational Survey Program, the Soldier's Manual, the RCA study of prerequisite competencies using TRADOC sources). In developing both training and MOS-specific performance measures, it will be desirable to maintain a file of these tasks for at least the MOS selected for special data collection.

4.3. Updating Procedures

Formal updating of the LRDB will be carefully controlled by the LRDB manager. It is essential that this be an orderly process to protect the integrity of the data base. Consequently, the procedures for modifying the LRDB will be made available only to the data base administrator and to the ARI data base monitor. Other requests to use these procedures for creating/updating other (non-LRDB)

files will be evaluated on merit and granted only with the approval of the ARI monitor and the data base administrator.

In many instances, file updates will involved adding derived variables or indices. In the course of analyses, a large number of such variables will be added to workfiles. Where the general applicability of such variables is judged to warrent the increase in storage space, these variables will be added to the master data base.

The process of adding new data to the file will involve several steps. These steps are designed to minimize the need for further changes or corrections once the data become available. Insofar as possible, such changes will be strictly avoided so as to eliminate the need for rerunning significant numbers of analyses to reflect corrected data. The steps to be followed in updating the file include the following:

4.3.1. IDENTIFICATION AND ACQUISITION OF NEW DATA AND RELATED DOCUMENTATION

In the case of the acquisition of existing data, this step will be relatively simple. For new measures to be collected, however, the data base staff will expect to play

a more significant role in the design of the data collection instruments to facilitate data entry.

For data that are not now in machine-readable form, the DB staff will provide for data entry. Plans call for the use of the interactive entry/edit system (FORMSPEC) available at AIR's Washington Office. If similar software can be installed at NIH, we will switch to entering data directly into NIH.

4.3.2. LINKING IN RELATED DATA

A separate Link file will be maintained to facilitate the addition of new data. This file will contain basic identifying information (SSN, name, birthdate, primary MOS, race, sex) and pointers to (index numbers) records in each of the relevant relations (files). Each new dataset will be passed against the Link file. For each matching record, all identifying information will be deleted and replaced with the appropriate pointers. For initial nonmatches, a second attempt will be made to match to the Link File on the basis of secondary identifiers, including, if necessary, manual inspection of "close" matches. For many new data sources, there will be a number of cases that are not already in the Link file. Where it is desired that

such cases be kept, relevant information will be added to the Link file and the data will be retained. This will be the case only if a new relation is being established so that it should cause no problem with the index values stored in the Link file.

Once all links to existing data have been established, the existing data needed for editing will be pulled out of the data base and merged with the new dataset. It is expected that this merge will be accomplished using SAS, since the edit procedures are designed as a SAS application.

4.3.3. EDITING

The editing procedures are described in detail in Section 3. They will typically involve two passes. In the first pass, specifications to detect errors and improbable values will be developed and implemented. After inspecting the results of this editing pass, error resolution specifications will be developed and implemented as a second pass.

4.3.4 DOCUMENTATION

Following the completion of the editing process, the documentation of the new data will be accomplished. A central part of this activity will be establishing the codebooks including frequencies and descriptive statistics as described in Section 5 of this plan.

4.3.5. MERGING THE NEW DATA

After the documentation has been completed, reviewed, and revised as necessary, the new data will be formally merged into the DBMS and appropriate backup tapes will be created (using the RAPID UNLOAD procedure).

4.3.6. DISSEMINATION

The final step in the addition of new information to the LRDB will be to inform potential users of the availability of the data and the documentation for the data. This will be accomplished through the electronic bulletin board implemented as part of the project sign-on procedures and through mailing to a list of ARI and project staff designated to receive information on the data base. This mailing list will be established and reviewed by the Project Director and Principal Investigator and by senior

ARI staff assigned responsibility for monitoring this activity.

4.3.7. EXCEPTIONS

It is expected that there will necessarily be exceptions to this orderly process. The most common form of exception is when quick analyses are required even though the data have not been completely edited. In most cases such analyses can proceed with the completion of step 2 (linking) and run in parallel with the editing. In a few cases, it may be necessary to strip identifiers and proceed with a copy of the input data only. In any event, the establishment of an orderly process makes the exceptions clearer. If preliminary analyses are run, they will be designated as such and checked as needed once the full update process has been completed.

4.4. Access

Primary access to the LRDB will be through the SAS interface procedure, PROC RAPRD. The Task 1 staff are all experienced SAS users and plan to conduct most of the analytic investigations using the SAS package. SAS is also the package of choice because of its capacity for merging

and transforming data easily. We intend to further simplify access to the data by creating a WYLBUR Command Procedure that will take a file name and variable list and create the SAS set up to read the requested variables into SAS and attach all of the appropriate variable labels and formats (for value labelling). This procedure will greatly simplify authorized access to the data base and will also contain a log file to monitor such accesses. (As discussed below, in Section 6, we will also place a logging procedure within the catalogued procedure that accesses the data base, as a further control on access.)

After the first portions of the data base are loaded, we will conduct a small cost-analysis to determine the relative efficiency of using RAPID operations to join information stored in different relations in comparison to using SAS merge operations to accomplish the same objective. This issue is not of major concern since SAS will accomplish this objective with reasonable efficiency, but it is of interest in cases where access from other procedures is required or where very large datasets are being created.

In addition to providing access to the data base through SAS, we will implement procedures for generating

SPSS systems files and also raw data files. A TPL interface is available, and we will install it if any need becomes apparent.

In general it is expected that requests for analysis files will be channelled through either the Project's Database Coordinator or ARI's monitor for this activity (who may also then pass the request on to the Database Coordinator). This pattern is expected to act both as a means of assuring a close monitoring of access to the data and also to insure reasonable efficiency since the data base staff will be most knowledgeable about the data base contents and access procedures. Except in very high priority cases, it is expected that workfile creation runs will be created overnight during the discount period. With the WYLBUR command procedures in place, we expect that workfiles can be created with only a 24-hour turnaround in most cases.

5. DATA DOCUMENTATION AND DISSEMINATION

5.1. Documentation Formats and Standards

Because the project involves the simultaneous collection and analysis of many interrelated sets of data by different teams of researchers, it demands particular effort in clear and complete documentation of the data base. This effort is complicated by the fact that the data base will not be constant, but rather will grow throughout the project as new measures are developed and new data are collected. It is essential, therefore, that the system for data documentation be carefully developed and strictly enforced from the outset of the project.

We will implement a multilevel system of interrelated data documentation documents that together will allow users easily to gain complete information on the data they may need to use. The key elements of this "metasystem" of documentation include:

- An Event File that documents each data collection "event";
- An Instrument File that contains copies of the data collection instruments (including

both questionnaires/tests and answer sheets where separate;

- A Sample Structure File giving a list of the different samples used in the data base and showing their relationship in Venn Diagram type format;

- A Dataset Log that shows the name, characteristics, and location of each data set in the data base and refers to the appropriate codebook documentation of the data set;

- SAS Codebooks for each data set, including frequency distributions for each discrete variable and complete summary statistics for continuous variables; for derived variables, the computational formula will be indicated; for other variables, the source file will be shown;

- Variable Cross-Reference Files listing each of the variables in the data base topically and by variable name and giving a list of all of the data sets for which the variable is available;

- o Data History Documentation for each data set, including an overall flowchart showing the steps and workfiles in the file creation/editing process and the printed output from each step in this process.

Each of these logs or files is described more fully below. We plan to use the WYLBUR text entry/editing capabilities to maintain "on-line" versions of all but the instrument file (where only the contents and index will be on-line). Hardcopy versions of each of these text files, as well as the instrument file, will also be maintained to facilitate the production and distribution of new copies of the complete documentation package for staff and others requiring such documentation.

5.1.1. EVENT FILE

The event file gives the basic "who, what, when, why, where" of each data collection effort. Specifically, each entry will include:

1. The date(s) and place(s) of the data collection events (e.g., FY81, at all MEPS; or June 23, 1983, at Fort Knox);

2. The sample(s) from whom data were obtained including the identifier used to access the Sample Structure File;

3. The instrument(s) used (including the instrument "identifier" used in accessing the Instrument File); and

4. A concise description of the purpose and intended use of the data. (For data collected by project staff, this will be a summary of the justification statement developed prior to the data collection and will refer to the more complete statement.)

5.1.2. INSTRUMENT FILE

Researchers occasionally need access to the original data collection instruments for such purposes as checking the actual wording of particular questions, checking potential skip patterns, and generating hypotheses concerning oddities in the responses. In many systems of data documentation, codebooks are constrained by variable and option labelling that must fit the format of the particular system employed. As a result, the full text of the question of the response alternatives is not available to the analyst. In addition, the "context" of the question

is not apparent in most codebooks. We will maintain a complete file of all of the instruments used, organized by an instrument identifier and accessed by instrument name and by a topical index of the instruments. Security restrictions may apply to some test instruments (e.g., ASVAB forms). We will investigate ways of satisfying security concerns in such cases.

The instrument file will be maintained in hardcopy form suitable for efficient copying on a Xerox 9400 as copies are needed for new project staff members or other researchers.

5.1.3. SAMPLE STRUCTURE FILE

Each time a new data set is received, the sample (and subsamples where appropriate) on which the data are based will be identified and assigned a sample identifier. This identifier, together with a more complete labelling of the sample(s), will be entered into the Sample Structure File that logs each sample and points to the relevant data set(s). The degree of overlap with every other sample will be ascertained and recorded in a sample structure matrix.

5.1.4. DATASET LOGS

Each of the raw and SAS data files comprising the data base will be listed in a data set log. This log will show all versions or generations of each data set beginning with the initial tape(s) or card(s) received from the field or from the data entry vendor. For each data set, the location (e.g., NIH tape library, NIH disk pack, backup facility tape library) will be indicated along with the primary data set characteristics (storage mode, block size, and record size where appropriate) and pointers to relevant entries in the Sample and Instrument Files. Much of this information will be maintained in the operating system's on-line catalog for the current, operative version of each file.

5.1.5. CODEBOOKS

Detailed information on each variable in the data base will be provided in SAS codebooks. The codebooks will be organized by dataset (relation) and data collection instrument with file and instrument identifiers indicated on the heading of each page. The specific contents of the codebooks will include:

1. a variable name and a more complete description for each variable;

2. a summary of the characteristics of the variable (character or numeric, number of characters/number of decimal places);

3. the number of cases with valid responses and the number for which the variable was omitted or missing;

4. a label for each response option for all discrete variables; and

5. the actual frequency distribution for each discrete variable and appropriate summary statistics (mean, standard deviation, median, quartile points, minimum and maximum) for each continuous variable.

5.1.6. VARIABLE CROSS-REFERENCE FILE

The Variable Cross-Reference File will contain an alphabetic and a topical listing of all of the variables in the entire data base. For each variable, the appropriate instruments, data sets, and samples will be indicated. The topical index will be of particular importance in providing

researchers a means of getting an efficient overview of the system and determining the availability of data to meet specific needs. During the planning phase, an initial variable taxonomy will be developed, and this taxonomy will be expanded during the project as appropriate. In developing this taxonomy, multiple listings for variables will be assumed (e.g., initial ASVAB scores might be listed under "Accession Data," under "Aptitude Measures," and also under "Performance Predictors").

5.1.7. DATA HISTORY DOCUMENTATION

Data history documentation will make it possible to examine each step in the creation and editing of the final datasets. This history documentation of each dataset will consist of a flowchart showing the files and programs used at each step in the creation/editing process and the output from each computer run in this process. The output will show both all of the program statements used and any printed results (e.g., record counts or warning messages). This documentation will be maintained on-line while the datasets are active.

5.2. Dissemination

Several different methods will be employed to make information on the data base available to appropriate individuals. News of immediate importance will be placed in an on-line electronic bulletin board with headlines announced through each user's logon profile. Similar on-line aids, accessible from Project accounts, will be used to point to the WYLBUR versions of the data documentation described in Section 5.1.

As the data from each new data collection become available, an informal workshop will be held. Printed copies of the documentation will be distributed to authorized users, and special characteristics of the data will be discussed. An initial workshop, held in May 1983, covered data storage and access procedures and information on RAPID, WYLBUR, and NIH computer facilities in general, as well as the detailed contents of the FY81/82 cohort files.

6. DATABASE SECURITY

6.1. The Need for Security

Whenever a large amount of data on individuals is maintained and stored, it is necessary to develop procedures to protect that data from compromise. The security of the Project A and B LRDB is particularly important for a number of reasons. Some of the data collected on individual soldiers, such as promotions, paygrade, or disciplinary actions, will be private in nature, and the privacy of that information must be maintained. Since many researchers will be accessing the LRDB for a variety of uses, the integrity of the data must be maintained to insure that the data remain accurate and consistent across uses. Finally, it is necessary to secure the data base to insure that the Army maintains ownership of the data. In other words, to insure that the data within the LRDB are used only for authorized Project A and B research.

6.2. Security Procedures

The security of the LRDB will be protected in a number of ways. Soldier social security numbers (SSN) will be routinely encrypted to insure the privacy of each soldier's records. Access to the LRDB will be controlled both to further protect soldier privacy and to insure proper use of the data. To provide further physical security, a log will be maintained for the LRDB system that will note each attempted access of the LRDB and whether the access was authorized or not. Finally, a set of data processing practices will be established to provide security for the information managing aspects of data within the LRDB. Each of these procedures will be detailed in the subsections that follow.

6.2.1. SSN ENCRYPTION

The key aspect to guaranteeing the privacy of individual soldier data will be the coding or encrypting of each soldier's identifier. This encryption will be accomplished by scrambling each soldiers' SSN in an unpredictable way. The algorithm that will do the encrypting (and if needed, decrypting) will be known only to the LRDB manager and ARI in-house data base

administrators. A printed copy of the algorithm will be securely maintained by the Project A COTR. All of the data files of the LRDB that can be routinely accessed and any project workfiles generated from the larger LRDB files will use only this encrypted SSN as the soldier identifier.

6.2.2. CONTROLLED FILE ACCESS

The integrity and accuracy of the LRDB data will be maintained by controlling the access to the large files or relations within the data base. This procedure will also further contribute to the privacy protection of individual soldier records. In general, the system to be adopted will use the RACF procedure available at NIH to allow the access of particular files to authorized users. Under RACF, different levels of access can be granted to different users. By specifying a "universal access" of "NONE," access can be restricted to only those users granted specific exceptions. In most cases, such users will be allowed "READ" access only. Such users will have to provide an eight-character RACF password (different for each user) in order to read the datafiles for which they have been given access. Using the provisions of RACF, a series of access "levels" will be developed which should provide timely access to relevant data needed by Project A

and B researchers and yet protect the security and integrity of the data.

Level 1. At the highest level of access will be the data base administrators. Currently these individuals are Dr. Lauress Wise and Ms. Winnie Young of AIR and Dr. Paul Rossmeissl and Ms. Frances Grafton of ARI. Level 1 personnel will have access to all of the files and relations within the data base. Furthermore, only Level 1 personnel will be able to enter data into the data base or modify data already stored in the data base. Thus, the data base administrators must assume responsibility for data entry, editing, and the storage of original data materials (i.e., tapes, punched cards) in a secure location. In addition, it will be the duty of the Level 1 personnel to create Level 4 workfiles as they are needed by other project researchers.

Level 2. Personnel at the second access level will be able to directly read data from all of the files in the data base with the exception of the Link File (see Section 4.3.2), which will contain basic soldier identifying information. This exception is made to maintain soldier privacy. It is planned that two members of the Project A staff will have Level 1 access to the LRDB. Dr. Ming-mei

Wang, the deputy Task 1 leader for statistical analyses, will need to be able to quickly access all data files, since the validation analyses of Task 1 span all tasks and data sets collected within Project A. Dr. Lawrence Hanser, the Task 4 monitor, will also be provided with Level 2 access. Dr. Hanser has been associated with Projects A and B since their inception and will backup the ARI in-house Level 1 LRDB staff in insuring that ARI has complete access to the LRDB for in-house research.

Level 3. Most project personnel will have some Level 3 LRDB access. Researchers at this level will have direct access to all files that are generated by the particular tasks they are investigating. Furthermore, they will have direct access to the files created by other tasks that directly impact their work. For example, Task 2 researchers will have direct access to the task analysis data collected by Task 5 so that the new predictors that are developed will address areas of the criterion space not currently covered by ASVAB.

Level 4. The most common way in which project researchers will access the LRDB is through the creation of workfiles (see Section 4.4.). By requesting the creation of a work file, a researcher will be able to obtain data

from all of the large files in the data base except the Link File which will contain soldier identifying information and will always be kept private and secure. The key aspect of workfiles relevant to LRDB security is that the researcher will only receive the data that he or she requested and there will be a precise record of who requested what data. When a project scientist requires a workfile, he or she will submit a data request form to either the contractor or ARI data base administrators. This request form will ask:

(1) Who wants the data?

(2) What variables are needed?

(3) What sample is needed?

(4) Which LRDB file or files contain the data being requested?

(5) Why are the data needed?

(6) Will the data be downloaded to hardware other than the NIH computer facility?

(7) What will be done with the data after its current use is completed (i.e., file will be scratched or saved for future use)?

In addition, each data request form will remind the researcher seeking data that LRDB is the property of the Army to be used only for Project A and B research and that all publications, papers, and briefing charts based upon these data must be submitted to ARI for clearance before they are presented to the public.

Paper copies of the workfile request will be available to all Project A and B scientists, but it is expected that most researchers will make use of a request form that will be stored on-line at NIH and can be quickly sent to a data base administrators using WYLBUR electronic mail. It is expected that most work file requests will be filled using overnight runs at NIH (see Section 4.4.) and should be ready within twenty-four hours of the original request for data.

6.2.3. LRDB LOG

The procedure used to execute the RAPID data base management systems' data retrieval programs has been modified to log a record of each access or attempted access to the data base. This access log file will be reviewed weekly to assure that no inappropriate access has been attempted. In addition, the monthly accounting information

of each project user will be monitored for any indication of unauthorized access to the LRDB. These audit trails will serve as a second level of protection against unauthorized use of the data by anyone who manages to obtain the necessary RACF passwords. They will not directly prevent unauthorized access to the LRDB, but the threat of exposure should serve as a significant deterrent to attempts at unauthorized LRDB access. The log will also help the data base administrators decide which project files should be stored on disk rather than tape by providing information as to how frequently data are requested from any given file.

6.2.4. OTHER PHYSICAL SECURITY PROCEDURES

Much of the data that will be entered into the LRDB will come from existing Army sources, such as the EMF. Additional precautions beyond those mentioned above will be taken to secure the information on these data tapes. The key aspect of this additional security is to collect and store information from these sources only if it is essential to the goals of Project A and B. For example, with regard to the EMF, this LRDB plan indicates specifically which variables will be needed. Other variables, in particular, each soldier's location and

assigned unit, will not be acquired in any form. In addition to limiting the data elements to be stored, the number of soldiers for whom any data will be retained will be limited. As indicated in Section 3 above, the LRDB will not obtain and keep information on all active service personnel. Only data from personnel selected for Project A and B research will be maintained.

6.2.5. DATA BASE ENTRY AND EDITING SECURITY PROCEDURES

In addition to providing for the physical security of the data base, procedures have been established to maintain soldier privacy within the areas of data base information management. Included within the broad topic of information management are such specific areas as handling of raw data, maintenance of raw data forms, and procedures for dealing with processed data (such as printouts or written reports). This section presents the procedures that will be used to provide security during data entry and editing, while the following section presents procedures that will be followed to protect soldier privacy in the analysis and reporting of data.

Data entry. All forms for data collected in the field will be shipped to the data entry station at AIR in

sequentially numbered packages via certified mail. Within 24 hours of their receipt, an entry in a data entry log will be noted for each package. This log will be maintained on-line, but will be backed up by a hardcopy following each log update. The log will contain identification of each package received, the number and type of the documents included, and the current status (entry preparation, entry, verification, editing, or shredded) of the data.

Data editing. While the data is being edited, the data collection forms (the raw data) will be stored in a locked room at a site removed from any post where the individual responses should not be of interest to anyone. Data integrity of the new data will be insured through thorough editing of the data. This editing will include: complete verification of all entered data, a reconciliation of the resultant record counts against the initial document counts, and relational editing of all new data to appropriate existing sources (e.g., the SSNs and birth dates match the master link file). Once the data have been entered into the LRDB and completely edited, the AIR data base administrator will review the completeness of the data entry/editing process. In performing this review, he will consult with the task leader and ARI task monitor

responsible for the data collection. Following any further revisions resulting from this review and a final approval of the editing, a back-up copy of the the resulting datafiles will be created that does not contain any personnel identifier other than the encrypted SSN. This tape will be removed from the NIH facilities and stored at a separate location.

Once the data have been backed up onto tape, all of the input documents will be shredded. A final count will be made of the number of documents shredded and this count will be checked against the initial document counts and the data entry log. At the same time that the data input forms are destroyed, all printouts generated during the editing of the data will be reviewed. Edit Run summaries and other general information will be found together to form the detailed documents of the editing process. Any other printouts, including any with potentially identifying information, will be destroyed along with the input documents. Likewise, any computer workfiles containing possible identifying information (excluding the master link file), along with all summary files not needed as backup or documentation, will be deleted from the system and then overwritten. The ARI data base administrator (or someone he delegates) will oversee this entire process.

6.2.6. DATA ANALYSES AND REPORTING SECURITY PROCEDURES

All workfiles, printouts, and analyses produced by the project data base personnel will contain a header indicating that the products were based on personnel data that the product should therefore be handled in an appropriate manner. When researchers are finished with any data, they will be required to specify the disposition of all workfiles and computer printouts that were created during the analyses. If work is of a contuining nature, a list of the workfiles and printouts will be retained for verification at the final completion of the analyses. When all analyses are completed, the ARI reviewer (the data base administrator or the appropriate task monitor) will approve the contents of any workfiles or printed documents that are to be retained. The primary purpose of this review is to assure that no information that might be used to infer individual soldier identities is retained.

All reports, journal articles, and conference papers, based on Project A and Project B research must be cleared by ARI before publication. This clearance process is primarily concerned with the political and scientific sensitivity of the research and typically is composed of three levels of clearance (team chief or task monitor, tech

area, and research laboratory). In the case of reports based on LRDB data, these reviews will be expanded to assure that information is not included in the reports that might eventually be used to ascertain the identify of individual soldiers.

6.3. Summary

Any set of procedures designed to store data electronically needs to balance the ease with which data can be accessed against the security of the data base. The procedures presented in this section tend to favor the security aspect of this balance. The number of data files that most project researchers will be able to access directly will be quite limited. Furthermore, only the data base administrators will be able to add or modify data, and access the true soldier identifying information. However, efficient use and rapid creation of the workfiles should provide any project scientist with the data that he or she needs to perform the required research.